OES Research Integrity Framework

The Office of Evaluation Sciences (OES) draws on diverse scientific expertise to (1) design improvements to Federal programs and policies and (2) rigorously evaluate the effectiveness of these changes. We evaluate effectiveness by running randomized evaluations in which individuals or groups are assigned at random to one or more versions of a program or policy. In this way we develop strong evidence about what works and what doesn't work — evidence that our agency collaborators can rely on when making decisions about how to design and run their programs moving forward.

Results from our studies impact the lives of millions of Americans, and thus the quality of our work and reliability of our findings are of paramount importance. At OES we take this responsibility very seriously, and we have developed a Research Integrity Framework designed to ensure that our research is conducted to the highest social scientific standards.

Three Core Principles: Reproducibility, Transparency, and Reliability

When we talk about research integrity, we mean that our work should be:

- **Reliable** — When we report that a program modification had a certain impact, or that one modification was better than another, or that a modification was ineffective, our collaborators in agencies across the Federal government should be able to trust that these findings mean what they purport to mean — that the research leading to these findings was conducted to the highest standards, that our statements about statistical significance are clear and correct, and that the limitations on our findings are also clear. Bottom line: Policymakers and program designers should be able to act on our findings with confidence.
- **Reproducible** —We conduct our work in such a way that we can carefully verify our results. All OES evaluations go through an internal replication, and any discrepancies between the two independent analyses are addressed before the results are finalized.
- **Transparent** — We are committed to ensuring that researchers, agency collaborators, policymakers, and the public at large are able to learn from our work. We keep a public record of all evaluations fielded and publicize all of our findings (including null findings and those that run counter to our own prior expectations and goals).

The overview below outlines key components of our Research Integrity Framework, including the six internal gates used to ensure quality across all OES projects. For further information about our work and our methods, please visit our website. There you can also find the document templates that we use to implement several of the gates described here, including templates for our Project Design Documents and Analysis Plans.

Five Steps We Take to Ensure Research Integrity

Design for Statistical Power

In designing evaluations, we give particular attention to statistical power. Briefly, statistical power is a study's ability to correctly detect whether a program improvement was effective (assuming that it was indeed effective). Among other things, statistical power depends on the number of cases included in a study and the method by which they are assigned to different treatment conditions. If a study lacks sufficient statistical power, then there is a risk of ending up with a "false negative" result. A "false negative" would be a failure to detect that a program modification really was effective, and this can have repercussions for future program design and policy making. When we vet our study designs to ensure they are as strong as possible, we pay particular attention to whether the study will have adequate power for the policy-makers decision. To progress to the "field" stage, every study must have adequate power to detect meaningful, policy-relevant effects.

Code Review

The defining feature and great advantage of randomized evaluations is the random assignment of individuals or groups to treatment conditions. This is what enables us to conclude that improvements in outcomes were actually *caused* by the policy or program changes that we tested. We generally use computer code to perform random assignment, and yet computer code is complex and notoriously vulnerable to mistakes. Before we use code for random assignment, we make sure it has been independently reviewed by an OES team member who is not directly involved in the project and thus has "fresh eyes." The reviewer works through the code line by line and may also test some or all of the code by running it on either real or mock data. By checking that our random assignment code is correct, we ensure that our agency collaborator's investment in a field evaluation is well founded and that, at the end of the project, the results mean what they are supposed to mean.

Analysis Plan Commitment

One of the most important steps we take is committing to a detailed analysis plan before we begin working with data. Why is this important? As the recent replication crisis in the social sciences has shown, if researchers allow themselves too much flexibility in analyzing data they may inadvertently get results that are the result of "fishing" or "*p*-hacking."[1] In our case, this would mean reporting "false positive" results that appear to indicate that a program or policy change was effective but instead reflect patterns or differences that appeared in the data by chance. To ensure that our positive results mean what they are supposed to mean, we commit to a detailed Analysis Plan before we analyze data — a best practice that has received greater attention in the social sciences in recent years. In particular, we commit ourselves to specific outcome variables

[1] See, for example: Gelman A, Loken E (2014) The statistical crisis in science. *American Scientist 102*(6): 460 ff.; Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22: 1359–1366.

and analytic methods, and we date-stamp the plan and post it on our website so that others can hold us accountable, other researchers can verify that our methods are sound, and policymakers can base decisions on our results with confidence.

At the end of a project, when we report our results and findings, we use the Analysis Plan to clearly distinguish between results based on planned (confirmatory) analyses and results based on unplanned (exploratory) analyses. In general, results based on planned analyses carry greater weight and provide strong evidence that a program or policy modification was effective in bringing about a change in outcomes. By contrast, results of unplanned analyses carry less weight; they should be treated as suggestive evidence and verified through further research. We are committed to drawing a strong distinction between these two types of evidence, and pre-committing to Analysis Plans is the principal way in which we do this.

Reanalysis

In keeping with our team's commitment to reproducibility, before we finalize an analysis, we submit it to an internal replication that we call Reanalysis. This is done by asking an independent reanalyst — an analyst who does not know the results of the initial analysis — to write new code to analyze the data and generate results that address the study's research objectives. Reanalysis serves as a check on (1) the computer code that the first analyst used to analyze the data, (2) any exploratory analyses that might have been conducted, and (3) any departures from the Analysis Plan that might have been necessary due to unanticipated features of the data.

The reanalyst's goal is to replicate the initial analysis from scratch, working only from the raw data and the Analysis Plan. It is important that the reanalyst not know the results of the initial analysis. Because we generally hold a team discussion of the initial analysis before Reanalysis occurs, we make sure that the reanalyst is recused from this discussion. Only when the reanalyst has finished do they look at the initial analyst's write-up of results and findings.

Publishing All Results

As part of our commitment to transparency and learning, OES shares all findings from every completed evaluation. This helps ensure federal partners can quickly learn what works and doesn't, and also learn from each others work. Results which are surprising or run counter to our expectations are just as important to share and often offer valuable lessons.

Our Project Process: Six Gates

To implement our Research Integrity Framework, we run our projects through a process that includes six **gates.** At each gate, the project is vetted against certain criteria before it can enter its next phase. Some of these gates emphasize general project management considerations such as feasibility, planning, and clear documentation, while others emphasize specific methodological

issues such as statistical power, pre-commitment to analysis plans, and reproducibility. The six gates are:

- **Project Initiation** - Each project is vetted early for feasibility, proper planning, and potential impact for stakeholders in a Federal program or policy.
- **Design Review** - Before any project can progress to the "field," phase, its intervention and evaluation design is carefully reviewed by select team members and then presented to the full team for comments.
- **Analysis Plan Commitment** - Before we work with data, we commit to an Analysis Plan and post it on our website.
- **Analysis Review** - An initial analysis of results is presented to the full team to check that the analysis is sound and comprehensive, that any limitations have been identified, and that alternative explanations have been addressed to the greatest extent possible.
- **Reanalysis** - After the initial analysis has been team-vetted and refined, it is checked by having an independent analyst, who is unaware of the initial results and findings, reproduce it. Discrepancies between the two independent analyses are then addressed before the results are finalized.
- **Pre-Publication Review -** To ensure transparency and reproducibility, all study materials are checked for completeness and proper archiving before a report is published.